

# Data Mining - Homework 1

Howie Benefiel *phb337*

February 18, 2019

## Problem 1.

(a) We start by calculating the sample standard deviation,  $\sigma_x = \frac{\text{sigma}}{\sqrt{n}} = \frac{1}{\sqrt{10000}} = 1e^{-2}$ . We then calculate the z-value,  $z = \frac{\bar{x} - \mu}{\sigma_x} = \frac{.1}{1e^{-2}} = 10$ . By consulting the Z table, we see that the probability that  $z_{avg} > .1$  is basically non-existent, it is less than .000001.

Repeating this process for .01, we get a z-score of 1. This corresponds to a 15.86% chance of  $z_{avg} > .01$ .

Finally, we see that  $P(z_{avg} > .001) = .4602$ .

(b) For the general case, we use the cumulative distribution function formula as follows:

$$\Phi(z) = \frac{1}{2} [1 + \text{erf}(\frac{z}{\sqrt{2}})] \quad (1)$$

We can then substitute in the formula for z to get:

$$\Phi(\frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}) = \frac{1}{2} [1 + \text{erf}(\frac{\frac{\bar{x} - \mu}{\sigma}}{\frac{1}{\sqrt{n}}})] = \frac{1}{2} [1 + \text{erf}(\frac{\bar{x} - \mu}{\sqrt{2}\sqrt{n}\sigma})] \quad (2)$$

Finally, to get the probability that  $z_{avg}$  is greater than some number we must subtract that formula from 1.

$$\Phi z = .5 - \text{erf}(\frac{\bar{x} - \mu}{\sqrt{2}\sqrt{n}\sigma}) \quad (3)$$

We can then plug in  $z_{avg}$ ,  $\mu$ , and  $n$ .

## Problem 2.

(a) We multiply out the operand of the sum to get

$$\frac{1}{n} \sum_{i=1}^n x_i^2 \beta^2 - 2x_i y_i \beta + y_i^2 \quad (4)$$

Since  $n$  ,  $x_i$  ,  $y_i$  are constants, we get

$$\min_{\beta} : A\beta^2 + B\beta + C$$

where

$$\begin{aligned} A &= \sum_{i=1}^n \frac{x_i^2}{n} \\ B &= \sum_{i=1}^n \frac{-2x_i y_i}{n} \\ C &= \sum_{i=1}^n \frac{y_i^2}{n} \end{aligned}$$

(b) From above,  $A = \sum_{i=1}^n \frac{x_i^2}{n}$  .  $n > 0$  because there are a strictly positive number of data points and  $x_i^2$  is always positive because any number squared is positive.

(c)

$$\begin{aligned} \min_{\beta} : A\beta^2 + B\beta + C &\iff 0 = \frac{d}{d\hat{\beta}}(A\hat{\beta}^2 + B\hat{\beta} + C) \\ 0 &= \frac{d}{d\hat{\beta}}A\hat{\beta}^2 + B\hat{\beta} + C \iff 0 = 2A\hat{\beta} + B \\ \hat{\beta} &= \frac{-B}{2A} \iff \hat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \end{aligned}$$

(d)

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

We then sub in  $y_i = \beta x_i + e_i$

$$\begin{aligned}
\hat{\beta} &= \frac{\sum_{i=1}^n x_i(x_i\beta + e_i)}{\sum_{i=1}^n x_i^2} \Leftrightarrow \\
\hat{\beta} &= \frac{\beta \sum_{i=1}^n x_i^2 + \sum_{i=1}^n x_i e_i}{\sum_{i=1}^n x_i^2} \Leftrightarrow \\
\hat{\beta} &= \beta + \frac{\sum_{i=1}^n x_i e_i}{\sum_{i=1}^n x_i^2} \Rightarrow \\
\mathbf{Z} &= \begin{bmatrix} \frac{x_1}{\sum_{i=1}^n x_i^2} \\ \frac{x_2}{\sum_{i=1}^n x_i^2} \\ \vdots \\ \frac{x_n}{\sum_{i=1}^n x_i^2} \end{bmatrix}
\end{aligned}$$